Statisztika

Politológus képzés

Daróczi Gergely

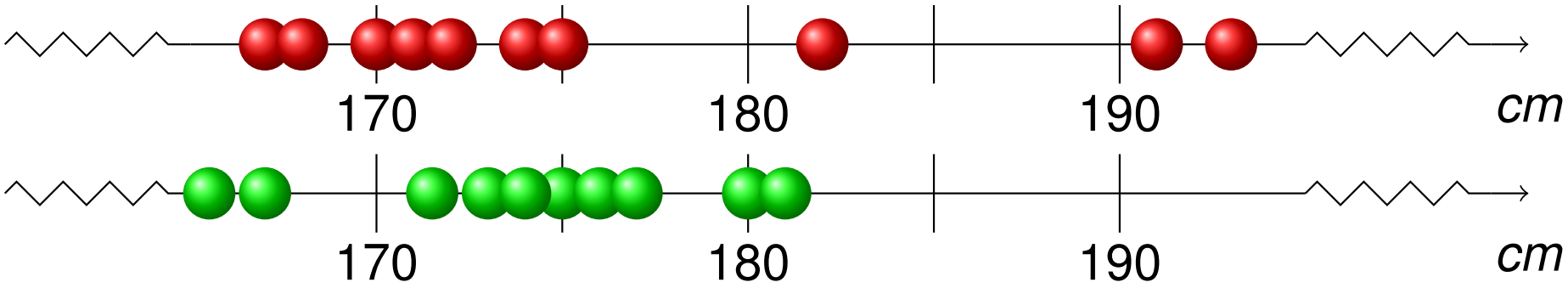
Politológia Tanszék

2011. április 27.

Pázmány Péter
Katolikus Egyetem
Bölcsészettudományi Kar

Outline

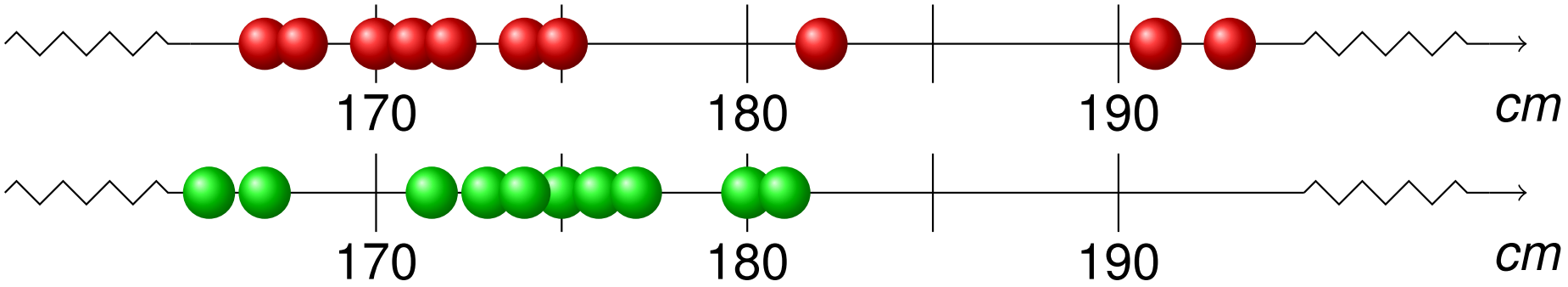
10-10 diák magasságát mértük 2 osztályteremben?

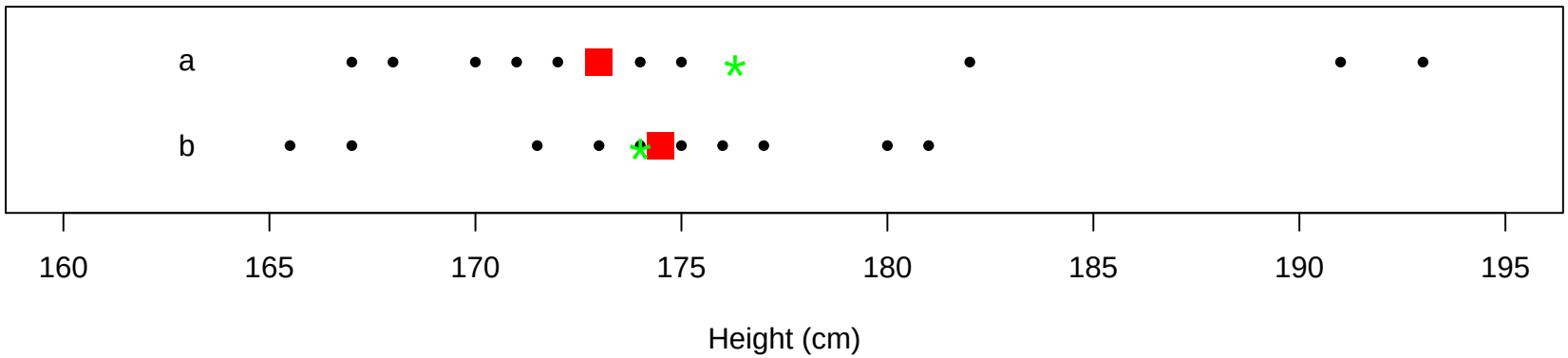


Melyik osztály diákjai a magassabbak a leíró statisztikák alapján adható becslések alapján?

Ismétlés

Középértékek





Általános iskolában végzett felmérés

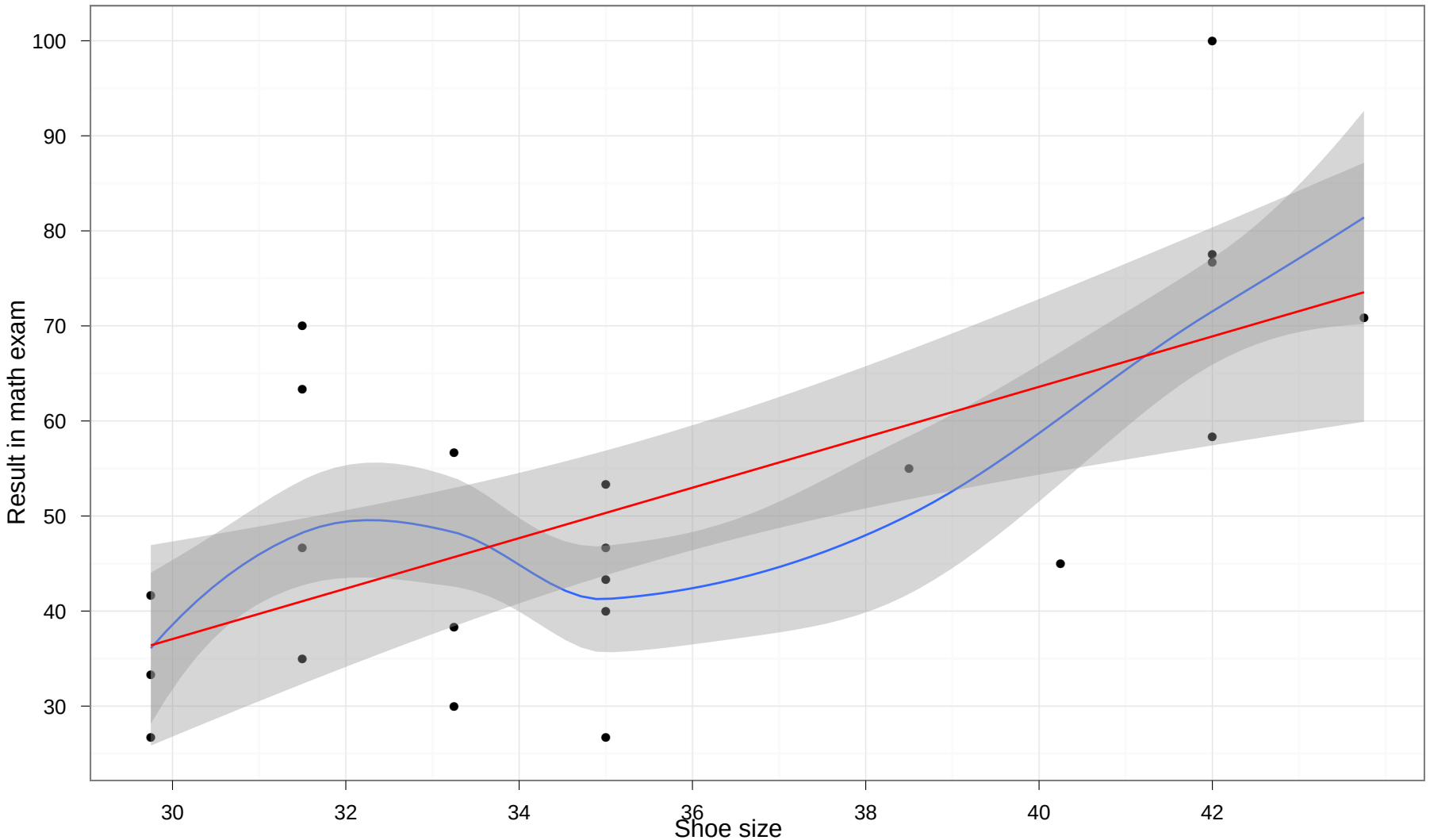
Cipőméret és IQ

Egy általános iskolában felmérést végeztünk a diákok cipőméretéről és egy matematika teszten nyújtott teljesítményükről. Az eredmények:

	Cipőméret	Matek	Kor
1	29.75	26.67	3
2	29.75	33.33	7
3	29.75	41.67	5
4	31.50	35.00	8
5	31.50	46.67	10
6	31.50	63.33	11
7	31.50	70.00	12
8	33.25	30.00	7.
9	33.25	38.33	7
10	33.25	56.67	12
11	35.00	26.67	6
12	35.00	40.00	8
13	35.00	43.33	6
14	35.00	46.67	10
15	35.00	53.33	11
16	38.50	55.00	9
17	40.25	45.00	9
18	42.00	58.33	9
19	42.00	76.67	16
20	42.00	77.50	18
21	42.00	100.00	19
22	43.75	70.83	14

Általános iskolában végzett felmérés

Cipőméret és IQ



Elméleti háttér

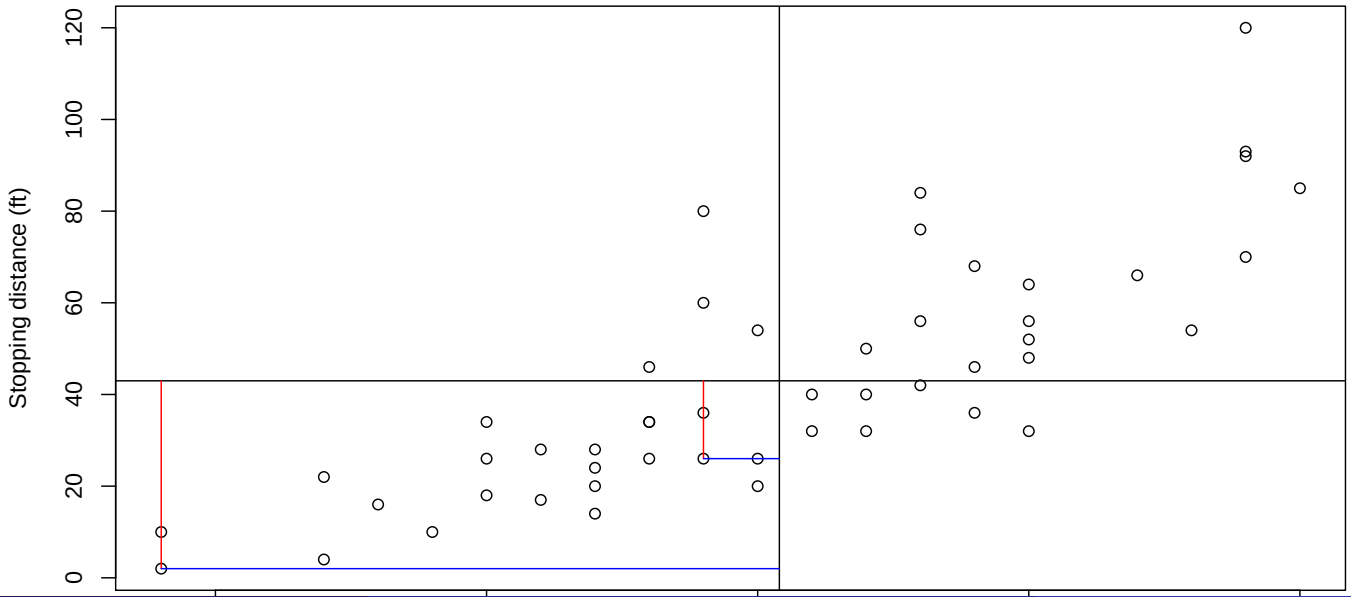
Covariation

x és *y* közös variabilitásának meghatározása:

$$COV(xy) = \sum_{i=1}^{n} \frac{(x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

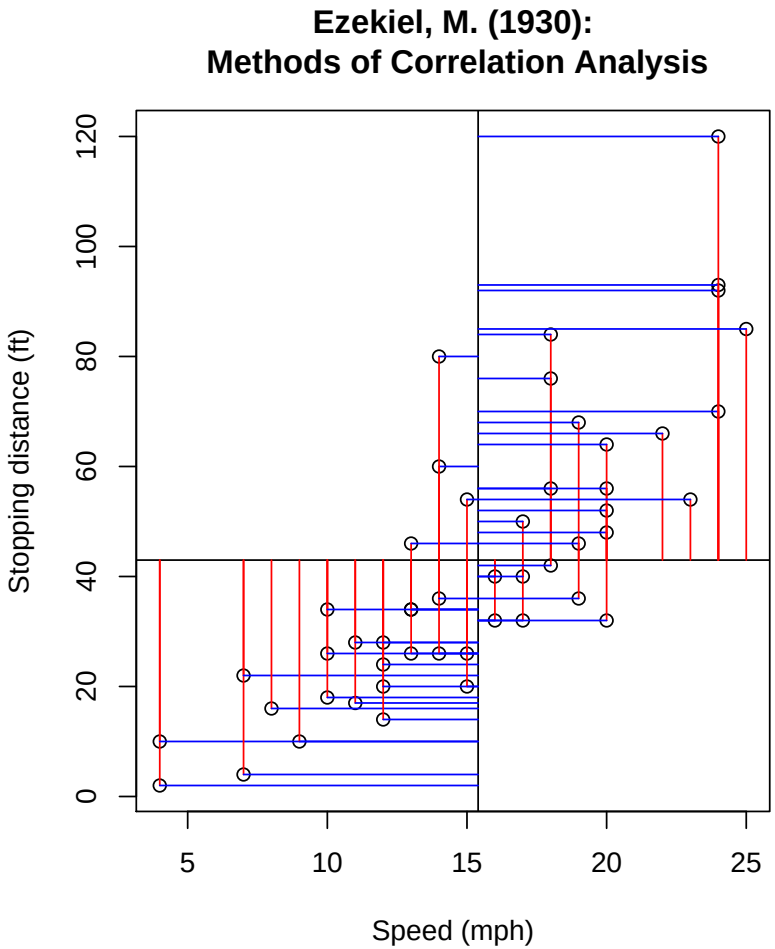
$$remember : \sigma = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n}}$$

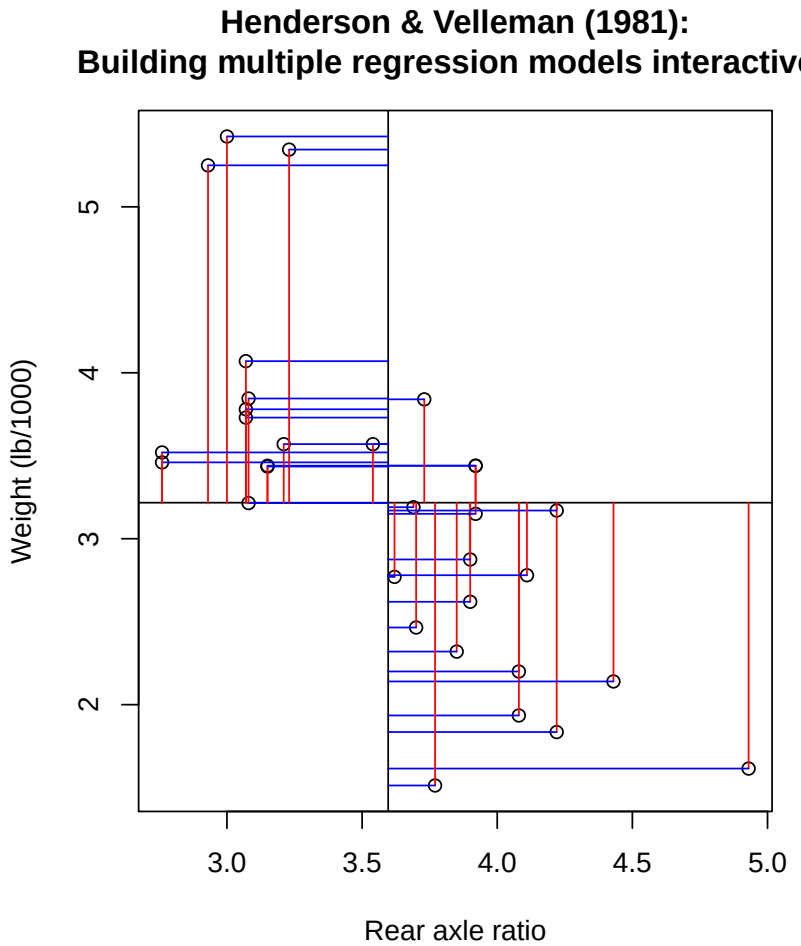
Ezekiel, M. (1930) Methods of Correlation Analysis. Wiley.



Elméleti háttér

Kovariancia

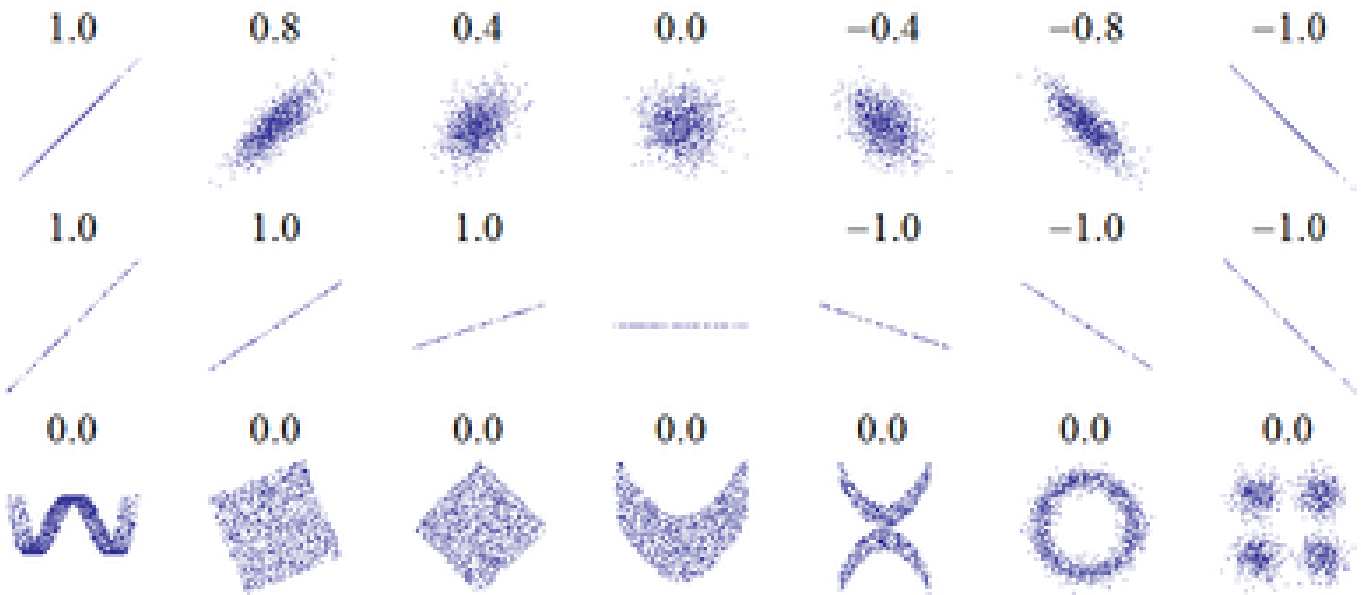




Elméleti háttér

Korreláció

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$



Elméleti háttér

Parciális korreláció

$$\hat{r}_{XY \cdot \mathbf{Z}} = \frac{N \sum_{i=1}^{N} r_{X,i} r_{Y,i} - \sum_{i=1}^{N} r_{X,i} \sum_{i=1}^{N} r_{Y,i}}{\sqrt{N \sum_{i=1}^{N} r_{X,i}^2 - \left(\sum_{i=1}^{N} r_{X,i} \right)^2} \sqrt{N \sum_{i=1}^{N} r_{Y,i}^2 - \left(\sum_{i=1}^{N} r_{Y,i} \right)^2}}$$

három változó esetében:

$$\hat{r}_{XY \cdot \mathbf{Z}} = \frac{r_{XY} - r_{X\mathbf{Z}} r_{Y\mathbf{Z}}}{\sqrt{(1 - r_{X\mathbf{Z}}^2)(1 - r_{Y\mathbf{Z}}^2)}}$$

Feladatok

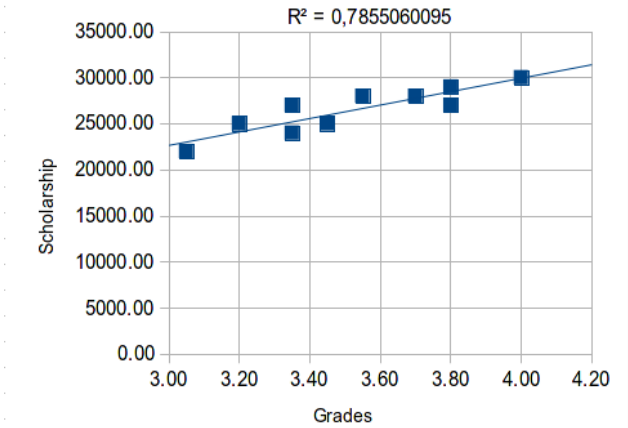
1. Határozd meg az alábbi adatbázis alapján a korrelációs együtthatókat!
2. Ne felejtkezz el a parciális tagokról!

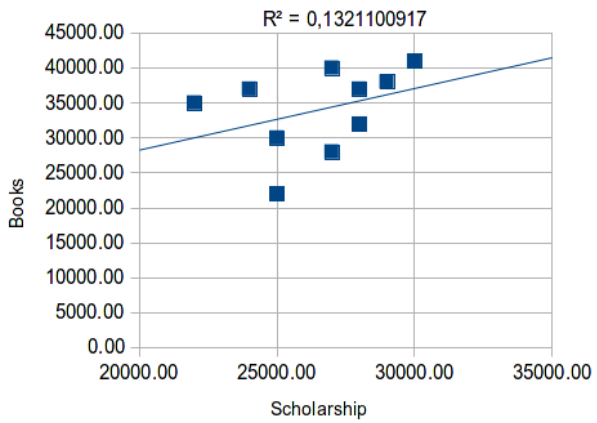
Tan. átlag	Ösztöndíj (HUF)	Könykiadás (HUF)
3.05	22000	3500
3.2	25000	3000
3.35	27000	2800
3.35	24000	3700
3.45	25000	2200
3.55	28000	3200
3.7	28000	3700
45	30000	4100
3.8	27000	4000
3.8	29000	3800

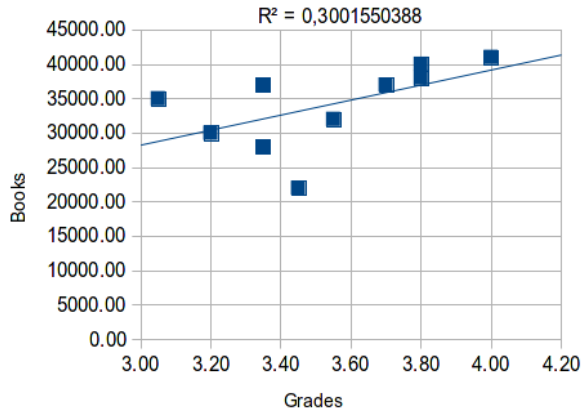
Feladatok

Megoldás

				$x_i - \bar{x}$			$(x_i - \bar{x})^2$			$(x_i - \bar{x})(y_i - \bar{y})$		
	Grade	Scholarship	Books	Grade	Scholarship	Books	Grade	Scholarship	Books	Grade-sch	Sch-books	Grade-books
	3.05	22000.00	35000.00	-0.48	-4500.00	1000.00	0.225625	20250000	1000000	2137.5	-4500000	-475
	3.20	25000.00	30000.00	-0.33	-1500.00	-4000.00	0.105625	2250000	16000000	487.5	6000000	1300
	3.35	27000.00	28000.00	-0.18	500.00	-6000.00	0.030625	250000	36000000	-87.5	-3000000	1050
	3.35	24000.00	37000.00	-0.18	-2500.00	3000.00	0.030625	6250000	9000000	437.5	-7500000	-525
	3.45	25000.00	22000.00	-0.07	-1500.00	-12000.00	0.005625	2250000	144000000	112.5	18000000	900
	3.55	28000.00	32000.00	0.02	1500.00	-2000.00	0.000625	2250000	4000000	37.5	-3000000	-50
	3.70	28000.00	37000.00	0.18	1500.00	3000.00	0.030625	2250000	9000000	262.5	4500000	525
	4.00	30000.00	41000.00	0.48	3500.00	7000.00	0.225625	12250000	49000000	1662.5	24500000	3325
	3.80	27000.00	40000.00	0.28	500.00	6000.00	0.075625	250000	36000000	137.5	3000000	1650
	3.80	29000.00	38000.00	0.28	2500.00	4000.00	0.075625	6250000	16000000	687.5	10000000	1100
Min	3.05	22000.00	22000.00			Sum	0.80625	54500000	320000000	5875	48000000	8800
Max	4.00	30000.00	41000.00			St. dev.	0.29930475	2460.80384	5962.84794	r	0.89	0.36	0.55
Range	0.95	8000.00	19000.00							r^2	0.79	0.13	0.30
Mean	3.53	26500.00	34000.00							partial corr	0.96	-0.24	0.46
Median	3.50	27000.00	36000.00										





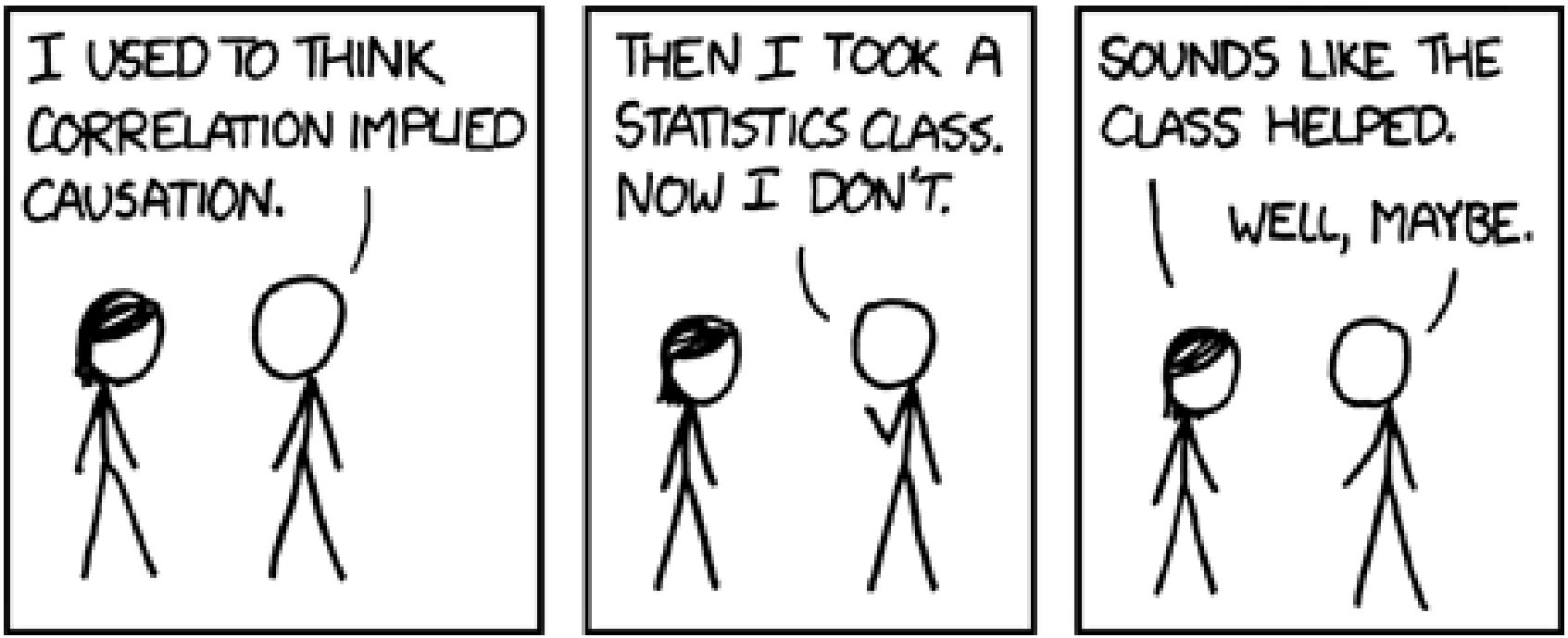


A korrelációs együttható korlátai

- Kauzalitás
- Lazarsfeld paradigma
- Linearitás

A korrelációs együttható korlátai

Correlation does not imply causation!



Source: http://xkcd.com/552

A korrelációs együttható korlátai

Correlation does not imply causation! - Elméleti háttér

Aristotle: logic, syllogism – if $(A \rightarrow B)\&(B \rightarrow C) \Rightarrow A \rightarrow C$

David Hume: scepticism

- „only correlation can actually be perceived [not causality]”
- see: our belief that the sun will rise tomorrow
- see: „If I see a billiard ball moving towards another, on a smooth table, I can easily conceive to stop upon contact.”

Popper: falsification

Pearl, J. - *Causality: Models, Reasoning, and Inference*, Cambridge University Press, 2000

Stouffer: *The American Soldier*

„Soldiers in branches with higher promotion rates are happier than soldiers in branches with lower rates of promotion.”

A korrelációs együttható korlátai

Lazarsfeld paradigma

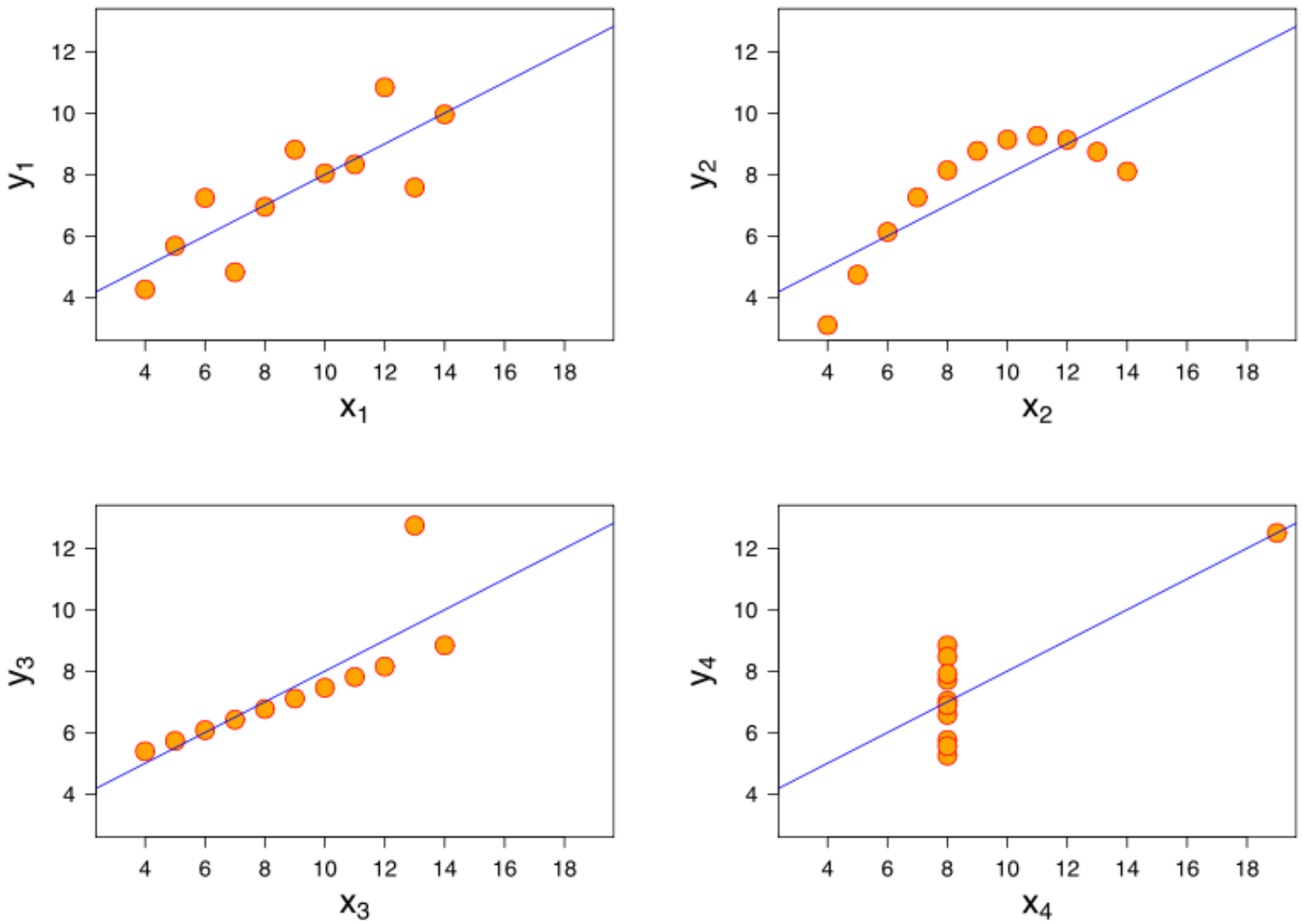
Stouffer: *The American Soldier*

H_0: Soldiers in branches with higher promotion rates are happier than soldiers in branches with lower rates of promotion. **BUT:**

„Soldiers in branches with higher promotion rates were more pessimistic about their own chances of being promoted than soldiers in branches with lower rates of promotion.”

A korrelációs együttható korlátai

Linearitás



Forrás: Anscombe, F. J. (1973) Graphs in statistical analysis. American Statistician,

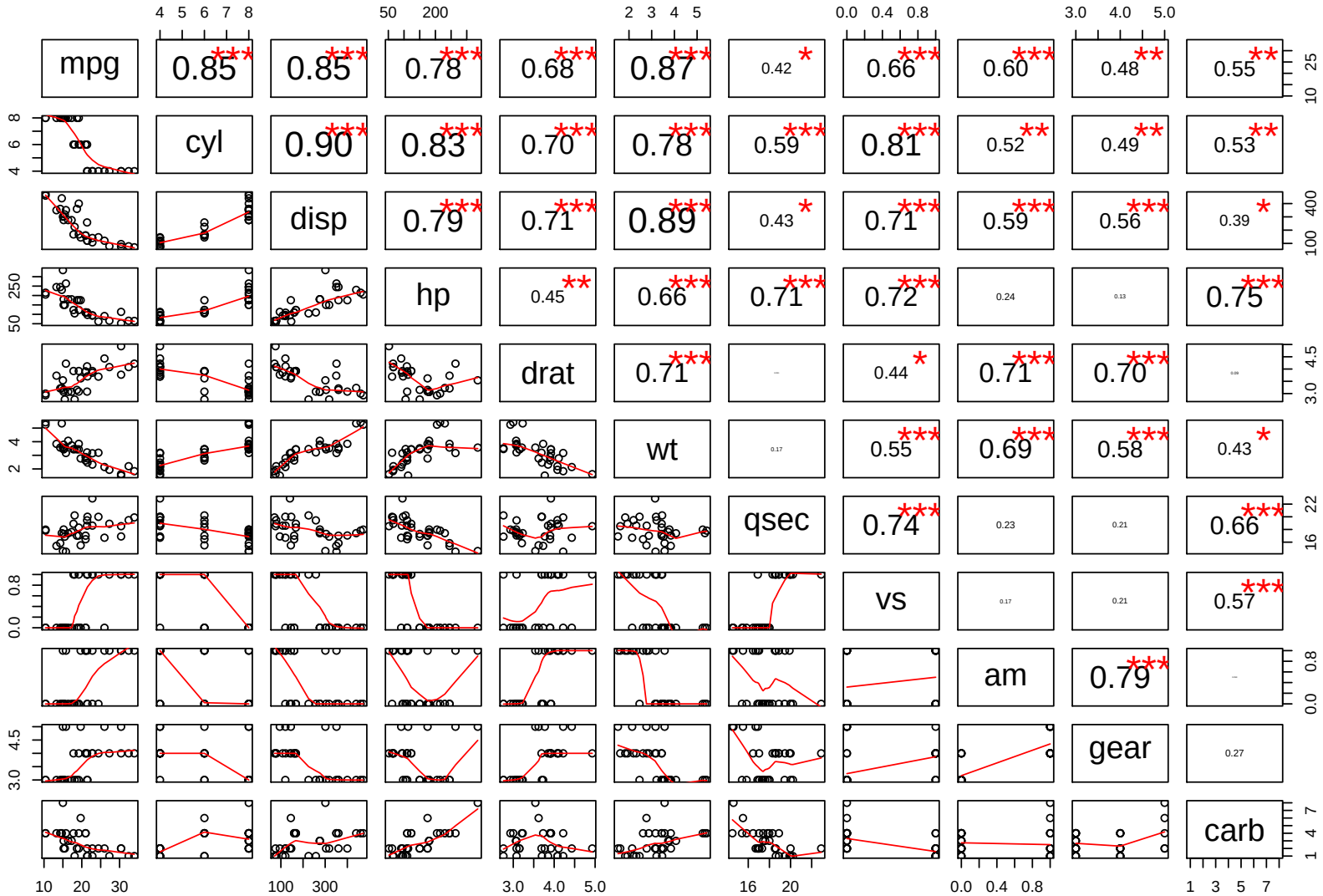
Feladat

The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models).

- mpg: Miles/(US) gallon
- cyl: Number of cylinders
- disp: Displacement (cu.in.)
- hp: Gross horsepower
- drat: Rear axle ratio
- wt: Weight (lb/1000)
- qsec: 1/4 mile time
- vs: V/S
- am: Transmission (0 = automatic, 1 = manual)
- gear: Number of forward gears
- carb: Number of carburetors

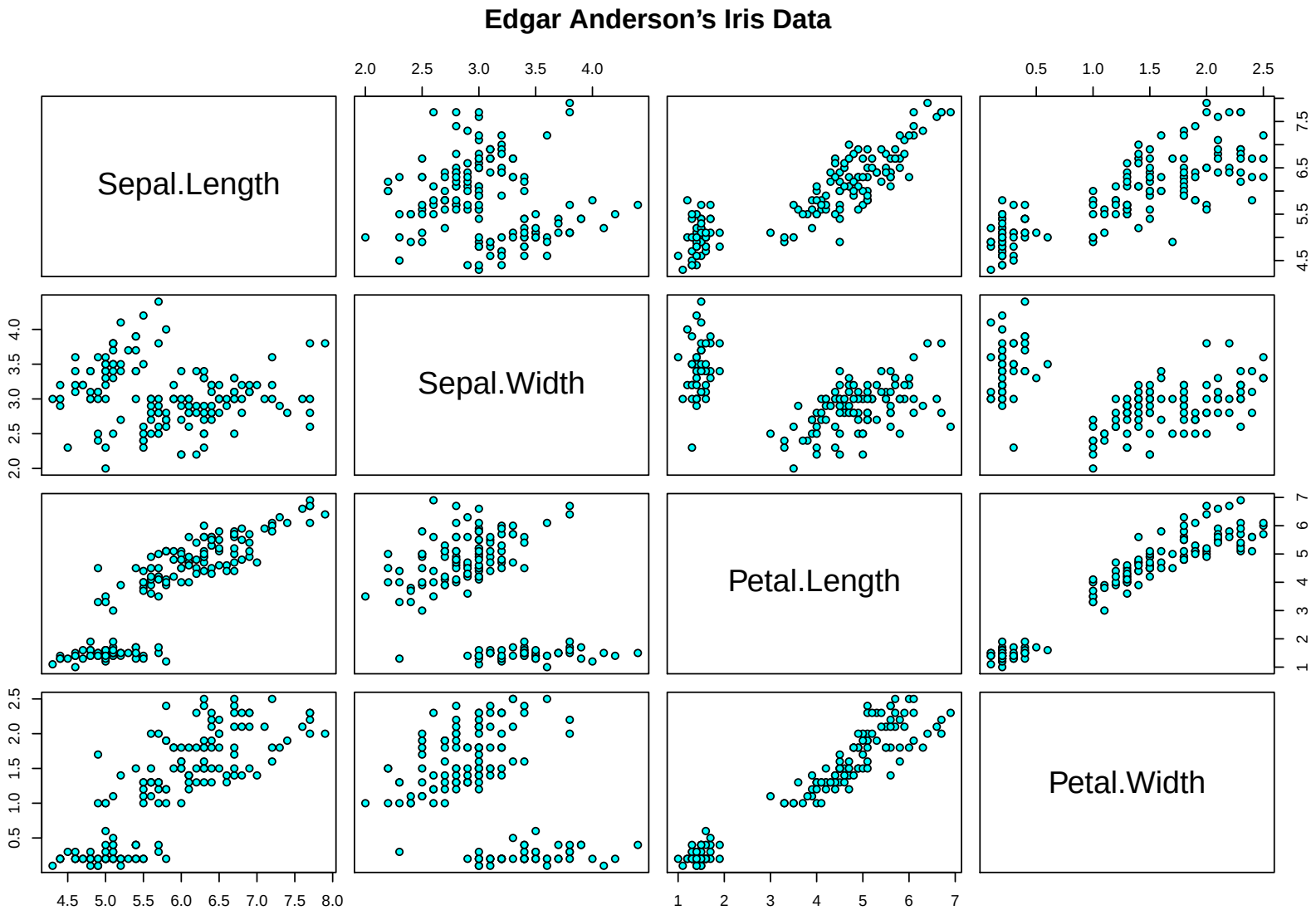
Source: Henderson and Velleman (1981), Building multiple regression models interactively. *Biometrics*, 37, 391-411.

Exercise



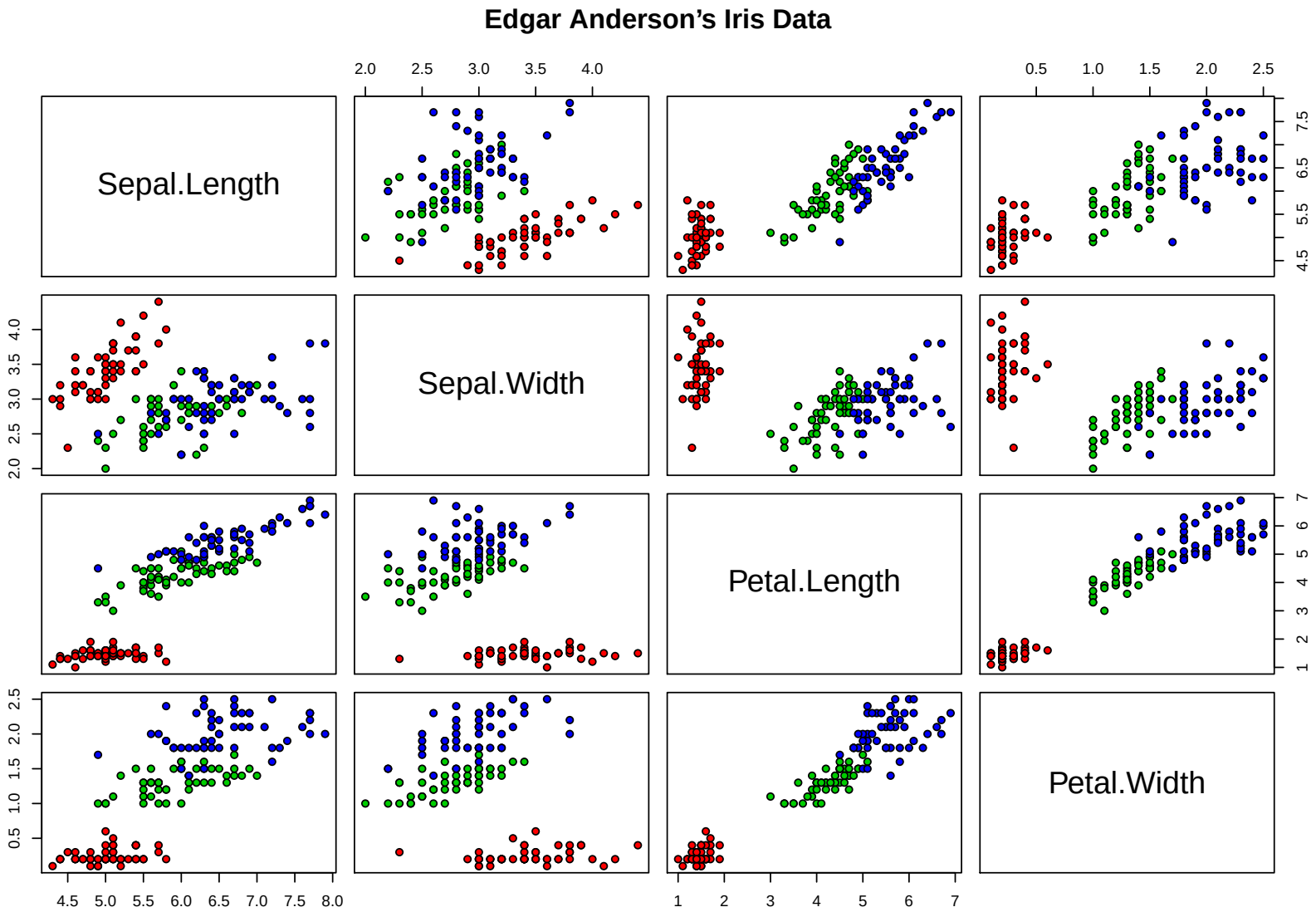
Henderson and Velleman (1981), Building multiple regression models interactively. *Biometrics*, 37, 391-411.

Exercise



Anderson, Edgar (1935). The irises of the Gaspe Peninsula, *Bulletin of the American Iris Society*, 59, 2-5.

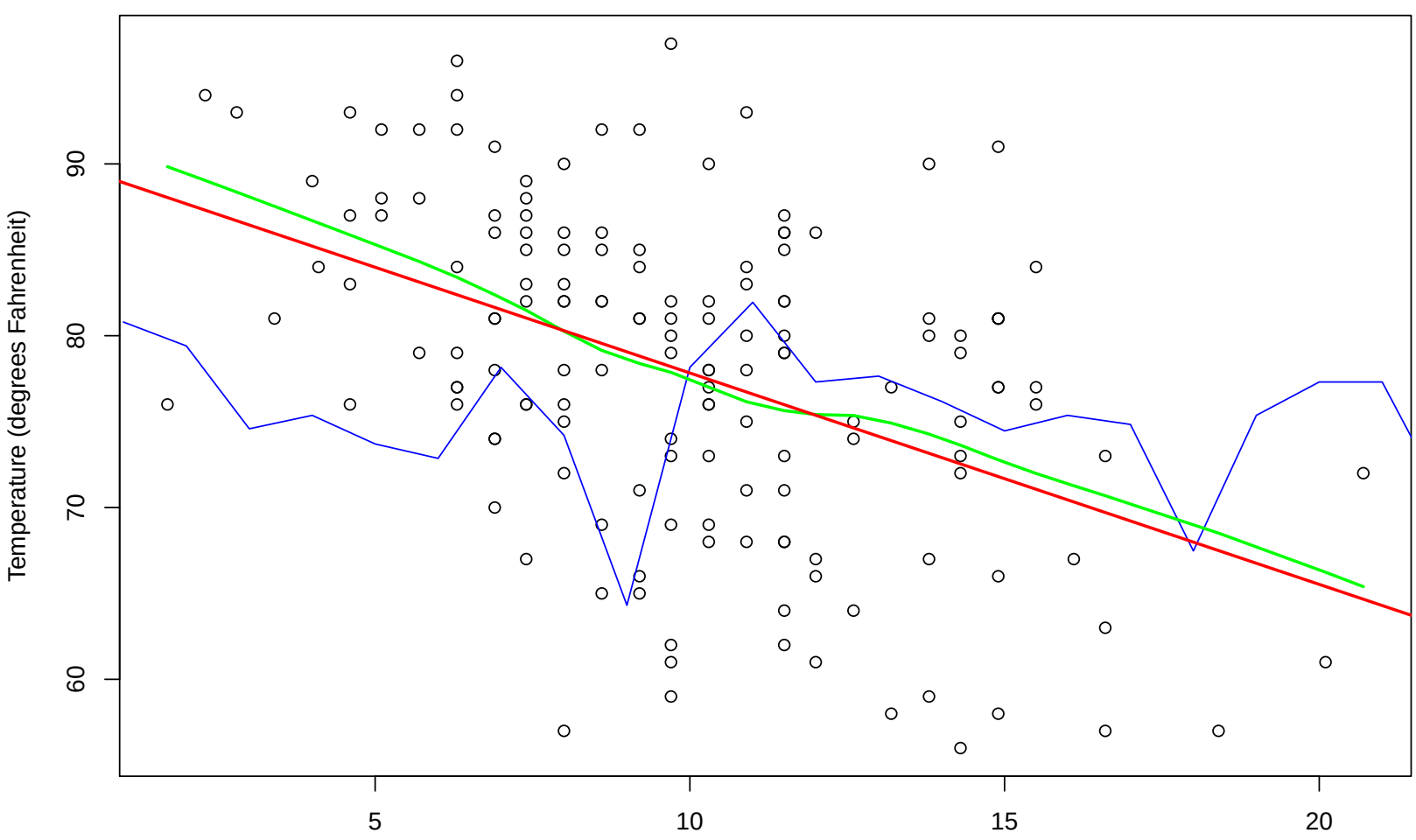
Exercise

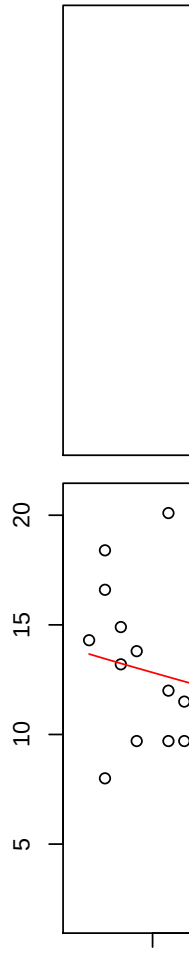


Anderson, Edgar (1935). The irises of the Gaspe Peninsula, *Bulletin of the American Iris Society*, 59, 2-5.

Újra korreláció

Valós kapcsolat?





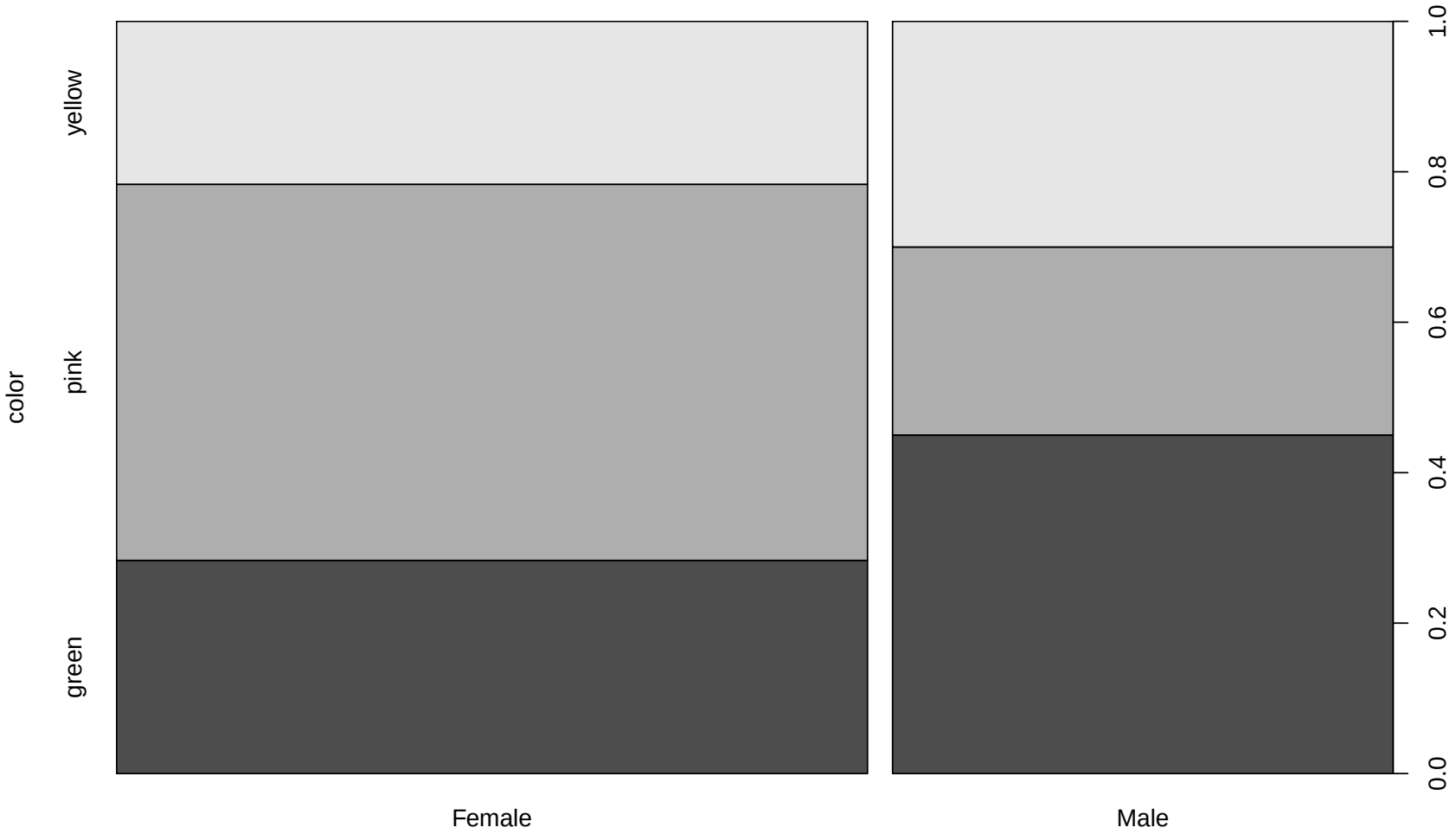
Kereszttáblák

Discrete (qualitative) variables

ID	gender	color
1	Female	pink
2	Female	pink
3	Female	pink
4	Female	pink
5	Female	pink
6	Female	pink
	...	
95	Male	yellow
96	Male	yellow
97	Male	yellow
98	Male	yellow
99	Male	yellow
100	Male	yellow

Crosstables

Discrete (qualitative) variables



Crosstables

Discrete (qualitative) variables

	green	pink	yellow
Female	17	30	13
Male	18	10	12

Crosstables

Discrete (qualitative) variables

	green	pink	yellow	
Female	17	30	13	2*Marginals
Male	18	10	12	
	Marginals			N

Crosstables

Discrete (qualitative) variables

	green	pink	yellow	$\sum$
Female	17	30	13	60
Male	18	10	12	40
$\sum$	40	35	25	100

Crosstables

Percentages

	green	pink	yellow	$\sum$
Female	17	30	13	60
Male	18	10	12	40
$\sum$	40	35	25	100

1. táblázat. Counted values

	green	pink	yellow	$\sum$
Female	17 %	30 %	13 %	60 %
Male	18 %	10 %	12 %	40 %
$\sum$	40 %	35 %	25 %	100 %

2. táblázat. Total percentages

Crosstables

Row percentages

	green	pink	yellow	$\sum$
Female	17	30	13	60
Male	18	10	12	40
$\sum$	40	35	25	100

3. táblázat. Counted values

	green	pink	yellow	$\sum$
Female	28.3 %	50 %	21.7 %	100 %
Male	45 %	25 %	30 %	100 %
$\sum$	35 %	40 %	25 %	100 %

4. táblázat. Row percentages

Crosstables

Column percentages

	green	pink	yellow	$\sum$
Female	17	30	13	60
Male	18	10	12	40
$\sum$	40	35	25	100

5. táblázat. Counted values

	green	pink	yellow	$\sum$
Female	48.63 %	75 %	52 %	60 %
Male	51.4 %	25 %	48 %	40 %
$\sum$	100 %	100 %	100 %	100 %

6. táblázat. Column percentages

Crosstables

Expected values

	green	pink	yellow	$\sum$
Female	17	30	13	60
Male	18	10	12	40
$\sum$	40	35	25	100

7. táblázat. Counted values

	green	pink	yellow	$\sum$
Female	21	24	15	60
Male	14	16	10	40
$\sum$	35	40	25	100

8. táblázat. Expected values

Crosstables

Chi-square statistic

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

where:

- χ^2: Pearson's cumulative test statistic,
- O_i: an observed (counted) frequency,
- E_i: an expected (theoretical) frequency,
- n: the number of cells in the table.

H_0: observed and expected values are all the same

Requirements!

Crosstables

Computed chi-square

	green	pink	yellow	$\sum$
Female	$\frac{(17-21)^2}{21}$	$\frac{(30-24)^2}{24}$	$\frac{(13-15)^2}{15}$	-
Male	$\frac{(18-14)^2}{14}$	$\frac{(10-16)^2}{16}$	$\frac{(12-10)^2}{10}$	-
$\sum$	-	-	-	-

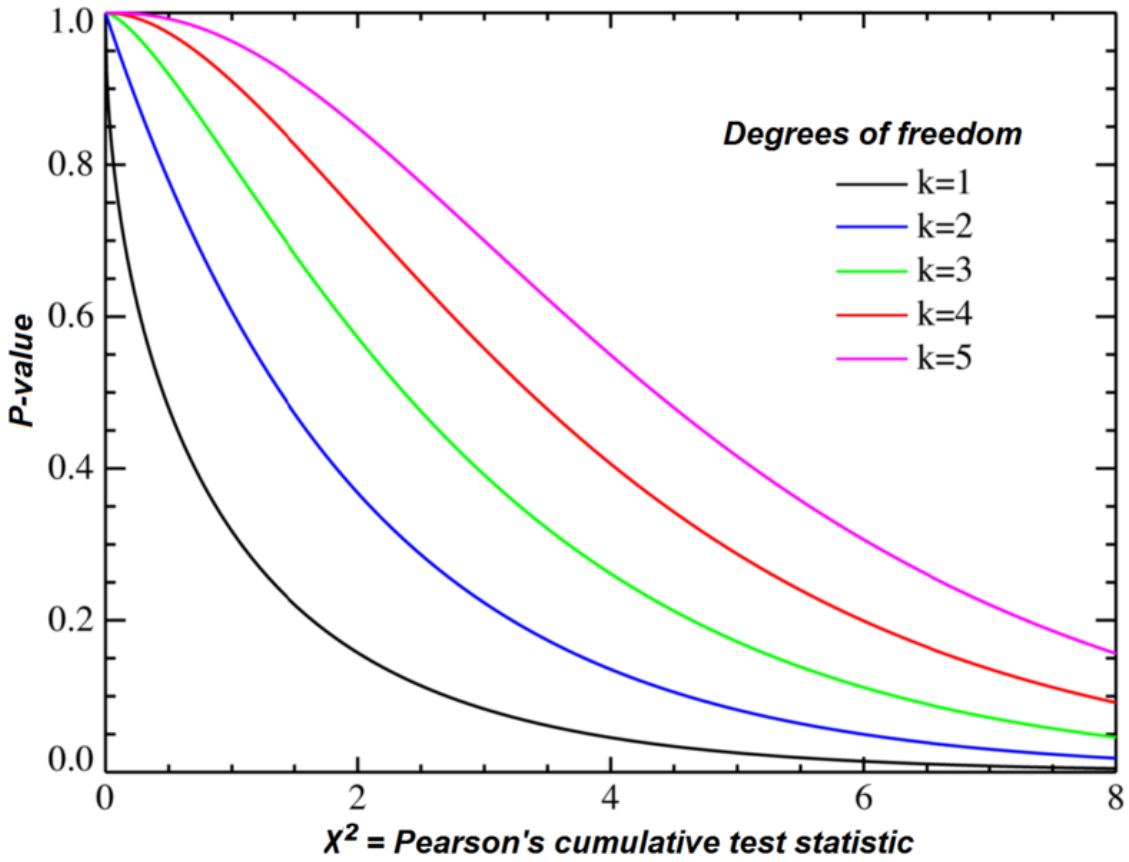
9. táblázat. Computed distances between observed and expected values

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = 6.321429$$

degrees of freedom: $(3-1)(2-1) = 2$

Crosstables

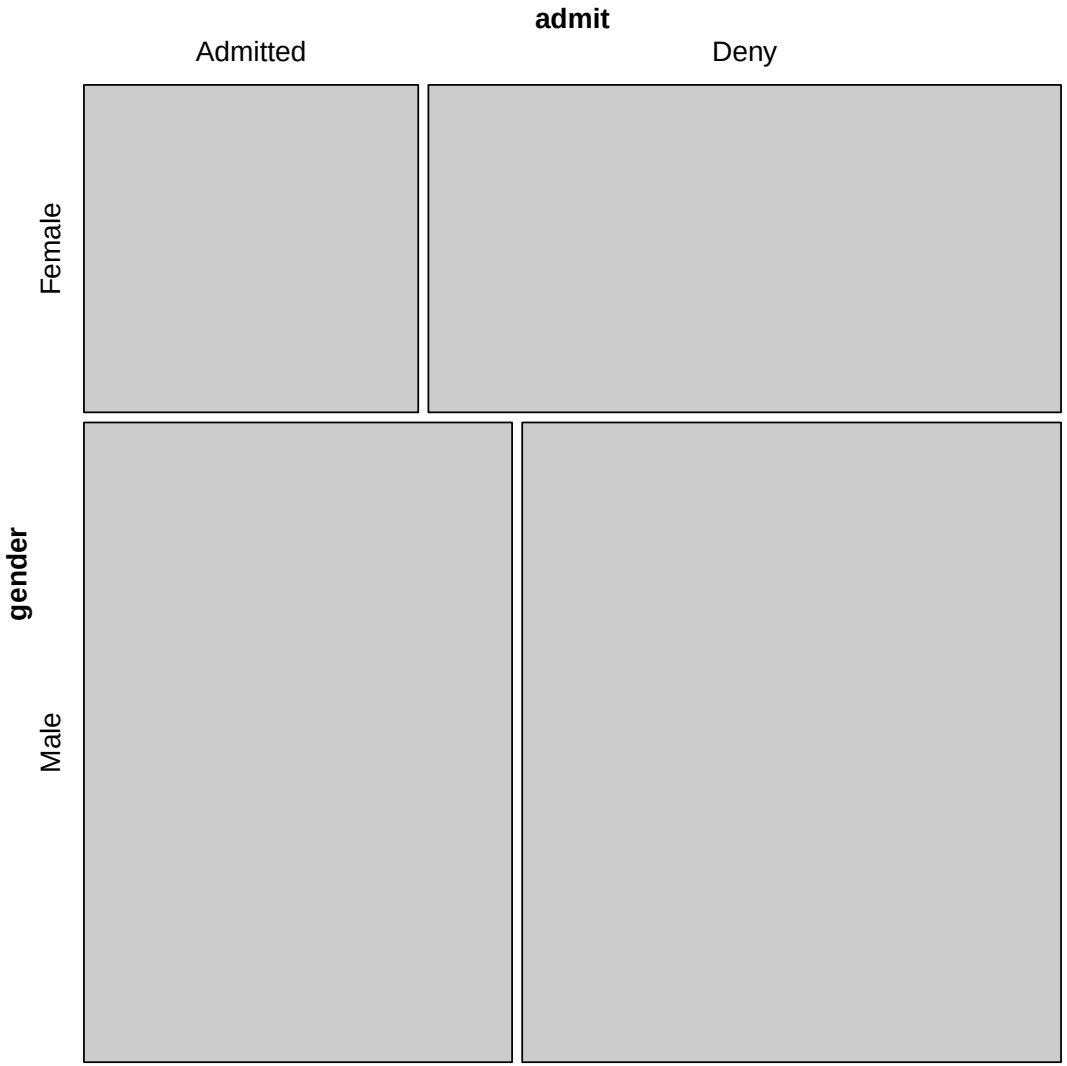
Computed chi-square



$$\Rightarrow p = 0.04239545$$

Simpson's paradox

Berkeley sex bias case



Simpson's paradox

Berkeley sex bias case

	Admitted	Deny	$\sum$
Female	1494	2827	4321
Male	3738	4704	8442
$\sum$	5232	7531	12763

10. táblázat. Observed values

	Admitted	Deny	$\sum$
Female	34.6 %	65.4 %	100 %
Male	44.3 %	55.7 %	100 %
$\sum$	41 %	59 %	100 %

11. táblázat. Row percentages

$$\chi^2 = 110.8489; d.f. = 1; p = 6.385628e-26$$

Simpson's paradox

Berkeley sex bias case

	Applicants	Admitted
Men	8442	**44%**
Women	4321	35%

	Men		Women	
Departement	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	272	6%	341	**7%**

Simpson's paradox

Batting averages in professional baseball

	1995		1996		Combined	
	Runs/Outs	%	Runs/Outs	%	Runs/Outs	%
Derek Jeter	12/48	25 %	183/582	31.4 %	195/630	**31** %
David Justice	104/411	**25.3** %	45/140	**32.1** %	149/551	27 %

Who is the better player?

To be continued. . .

Daróczi Gergely
daroczi.gergely@btk.ppke.hu